Hodnocení diplomové práce

Jméno studenta:	**Mgr. Jana Tíkalová**
Téma diplomové práce:	**Scoring a odhad bonity klienta**
Hodnotitel – oponent práce:	František Vávra
	Západočeská univerzita v Plzni
	Fakulta aplikovaných věd
	Katedra matematiky

Předložená diplomová práce se zabývá aktuální, často řešenou problematikou a ve své úplnosti zatím nedořešenou.

Hodnocení obsahové stránky

Vyjadřování v práci je povrchní a velice často více-významové. U statistických metod (zvl. logitu) postrádám alespoň některé předpoklady jejich užití. Obecně je statistické vyjadřování silně fuzzy (neurčité, nejasné, nepřesné). To platí nejen pro statistické výroky. Vzhledem k tomu, že se jedná (v aplikační části) o řešení zcela konkrétní úlohy, postrádám popsání a modelování chyb I. a II. druhu. Z textu jsem nenabyl přesvědčení, že v práci uváděné výsledky a některé vztahy jsou správné a prakticky využitelné.

Konkrétní připomínky (některé příklady):

- Str. 3, typologický přehled rizik je příliš učebnicový a tedy povrchní, pod likvidním rizikem je uvažováno i riziko solventnosti. To je ale podstatný rozdíl, solventní klient (v majetkovém pohledu) nemusí být likvidní! Nikoliv naopak!
- Str. 4, opět typologie, ideální pro povrchní popis, nikoliv pro algoritmické a analytické využití.
- Str. 7, odkaz na „Vyhlášku [16] vede v seznamu literatury na něco zcela jiného, navíc v seznamu literatury je ponecháno „nedoplněné ISBN".
- Str. 7 a další, vymezení defaultu (kapitola 2.3.1) je pro statistické aplikace poněkud vágní a široké (viz i další text v aplikační části práce), zcela určitě (v případě testování hypotéz) povede na složenou hypotézu a složenou alternativu (viz str. 9), což je netriviální, těžko řešitelný problém.
- Obecně sub-kapitoly 2.3.1 – 2.3.4 jsou poplatné zvykovému rozmazanému vyjadřování, tedy vhodné pro popularizaci problematiky, nikoliv pro korektní statistickou inferenci. Navíc, evidentně se jedná o převzaté myšlenky s nepříliš jasným vyjádřením zdroje(ů).
- Str. 11, sub-kapitola 3.1. Zde evidentně chybí přehled elementárních požadavků na scoringovou funkci z pohledu jejího následného využívání. Score testu nějaké hypotézy proti nějaké alternativě může, ale nemusí (častější) být vhodným vyjádřením bonity klienta. Zde je též problém s pojmem „bonita klienta" ve zcela exaktním smyslu. Zdá se, z textu práce, že takových „bonit" může být více!
- Kapitola 3.2. Zde se nechá tušit, že se autorka v dané problematice přijatelně orientuje, leč vyjadřování zůstává na popularizační úrovni. Exaktnost chybí. Klasickým příkladem je typologie závislostí. Je smíchána dohromady „závislost jevů" a „závislost náhodných proměnných", a to bez přihlédnutí k jejich provázanosti. Pojmy jsou používány svévolně (deterministická, stochastická, statistická, ... závislost). Obsah str. 15-17 jsou klasické pravděpodobnostní a statistické vzorce, bez jasného rozlišení,

kdy se jedná o parametr rozdělení a kdy se jedná o jeho odhad. Chybí odkazy na zdroje proto např. vzorec-testové kritérium (3.5) je možno chápat jako univerzálně použitelné a to zcela určitě není!

- Str. 18. Tvrzení „Nejčastěji využívaný přístup k analýze dat různé povahy je regrese." je patrně pravdivé, leč regresní modely vyžadují splnění poměrně přísných předpokladů. Proto i jejich výsledky bývají problematické. A to vše i bez ohledu na velice svérázný názor autorky, že u logitové regrese je absence přísných předpokladů na vstupní proměnné. Taktéž poznámka pod čarou č. [18] je ve vyjádření, že cosi umí neparametrické metody (multikolinearita), velice sporná, autorka by je měla doložit příklady! Poznámka o nesnadné interpretaci neparametrických metod bankovními úředníky je vlastně popřením tématu diplomové práce. Scoring (popularizačně) není nic jiného než „překlad" statistických, případně i pravděpodobnostních, výsledků do jazyka a pojmů „bankovních úředníků"!
- Str. 20. Předpoklad **B,** vyjadřuje jen homoskedasticitu, pokud jde o „sériovou nezávislost" pak jen v případě, že náhodné složky $u_1, \ldots, u_n$ mají sdružené nedegenerované normální rozdělení! Předpoklad **D**: Matice **X** nemusí mít plnou hodnost i v případě, že každé její dva sloupce jsou nezávislé (mohou se vyskytovat lineárně závislé trojice, ...). Co to je „perfektně závislé sloupce"? Tvrzení, že vlastnosti **A-C** jsou ekvivalentní $u \sim N_n(0,\sigma^2\mathbf{I})$ je špatně! Platí splnění $u \sim N_n(0,\sigma^2\mathbf{I})$ garantuje **A-C**, nikoliv naopak!
- Str. 22. I bonitní klient se může dopustit defaultu (nechce platit, i když na to má). Poznámka [22] pod čarou fakticky říká, že proměnné u lineárního regresního modelu a u „logitového modelu" jsou vzájemně nepoužitelné. To je jasný spor s výrazem (3.22)!
- Str. 23. Odkud první věta na této straně? Odstavec *„Z výše uvedeného plyne, že po provedení logistické transformace závisle proměnné nejsou v modelu splněny předpoklady lineárně regresního modelu, proto nelze lineární regresi ani aplikovat na naše data. Logitová regrese tak, jak je popsaná v této práci, tyto předpoklady na data neklade a pro její použití není nutné ani ověřovat tyto předpoklady."* na konci úvodní sub-kapitoly kapitoly 3.5 je neprokázaným a nesmyslným dojmem autorky (jedná se opět o spor se vztahem (3.22)).
- Str. 24, 25, 26. U vztahů (3.27 a 3.28) chybí podmínky existence a jednoznačnosti jejich řešení! Tím i následná odvození jsou přinejmenším sporná!
- Str. 26, poslední odstavec sub-kapitoly 3.5.1 je bez dalšího upřesnění chybně (platí jen pro jednoduché hypotézy). V předchozím textu bylo „mnoho H_0", které jsou složené.
- Sub-kapitola 3.5.2. Jedná se o převzaté výsledky z různých zdrojů, kde nebylo upraveno individuální značení (a ani není jednoznačně uvedeno, co, to které značení znamená) z jednotlivých zdrojů. Tím dochází k sporným vyjádřením.
- Totéž platí i pro kapitolu 3.6.
- Kapitola 4. Jedná se o přiměřeně fundovaný popis disponibilních dat a cíle práce. I zde se však projevuje povrchnost zpracování Např. poslední odstavec v kapitole 4.1. (str. 37). Zde je popisována poměrně složitá struktura jevu zvaného „default" aniž by byla nějak zjevena vazba na text v sub kapitole 2.3.1. Případně i kriticky!
- V kapitole 5. je poněkolikáté diskutována problematika chybějících údajů v jednotlivých záznamech. Z textu je zřejmé, že autorka chápe závažnost tohoto jevu. Řešení však nalézá ve vynechávání záznamů. Tedy nejstupidnější varianta.
- Sub-kapitola 5.3.1 (str. 49). Není vůbec zřejmé, jak byla disponibilní data dělena (slovo „prvních" může mít, bez dalšího mnoho významů) na soubor pro tvorbu

modelu a soubor pro testování. To zcela znehodnocuje následující testové, ověřovací, výsledky! Viz i další poznámky.

- Str. 49. Nacházejí se zde dvě vyjádření:
  - Veškeré nezávislé proměnné[1] vystupující v našem modelu jsou kategoriálního typu (rozdělené do intervalů) a jejich kódování je dané společností, která nám data poskytla.
  - Na základě Pearsonova koeficientu korelace jsme určili stupeň korelovanosti mezi proměnnými, přičemž nejvyšší hodnotu .... .

  Zde, z těchto dvou výroků, jsou namístě dvě (možná více) zásadní otázky:
  - **Co to je korelace mezi kategoriálními náhodnými proměnnými?**
  - **Co si proto lze myslet o následně uvedených výsledcích?**
  - Nejde o korelaci mezi kódy jednotlivých kvant (příslušnost do intervalu)? Bylo takové případné kódování ordinální?
  - Proč pak Pearsonův koeficient korelace?
  - ... ?
- Str. 53. V dolním odstavci na této stránce je zmíněna problematičnost získaného výsledku „prý" volbou prahového odhadu pravděpodobnosti „defaultu". A to bez dalšího rozboru dokládajícího toto tvrzení.
- Str. 55. Zde uvedený hladový algoritmus „odstraňování nevýznamných proměnných" není opodstatněn (i když z uvedeného pramene doporučen) a mohl by mít i negativní důsledky. Ty možná naznačuje text posledního odstavce na této straně. Zde je v rozporu se vztahy (3.25 a 3.26) tvrzeno, že lepší model má nižší hodnotu logaritmu věrohodnosti. To je ale evidentně špatně!
- Str. 56, dole. Zde je uvedeno, že bylo 390 predikovaných jevů ohodnoceno dobře. To je ale poněkud podivné číslo, neboť k dispozici bylo celkem 769 pozorování (str. 48), z toho bylo pro statistickou inferenci použito 450 (str. 49). Odtud pro testování zbývá 319. **Kde a jak se mezi nimi (mezi 319) mohlo najít 390 správně testovaných.** Text také naznačuje problematiku diference chyb prvního a druhého druhu. Ty, v této konkrétní aplikaci, mají jinou povahu než v čistě statistické formulaci úlohy. Default klienta (není li předpovězen) může „stát" celou nebo část půjčky + povinně tvořené rezervy, zatímco neposkytnutá (ne-defaultní) půjčka bude stát „nevytvořený zisk po zdanění". Problém je zde také ve statistickém odhadu chyby II. druhu (nejen). Tuto informaci nelze korektně zjistit[2].
- Str. 56-57. Specielně srovnání mezi tabulkami 5.10 a 5.11. Je zde problém nestatisticky (bankovně, finančně, datově, ...) vysvětlit odstraněné a ponechané modelové proměnné (proč např. vypadla proměnná nazvaná ***„platební morálka"***, ...?).
- Str. 60. Zde a často i na stranách předchozích a i následujících se mluví o jakési „podmíněné pravděpodobnosti" defaultu. Jak se liší od deklarované $\pi(x)$?
- Str. 61 a následující. Zde se mluví o nějakém logitu vázaném na jednotlivé proměnné a je pak graficky testována „linearita" přes jednotlivé kvantovací intervaly. O co jde? Jedná se o dílčí složku logitového skóre? Nebo se jedná o průběh „celkového" logitového skóre v závislosti na změnách jedné proměnné, při pevných hodnotách ostatních proměnných (při jakých?)? Pokud by šlo o první variantu, pak grafický test nemá smysl (nepřímkový průběh jedné by mohl být kompenzován na přímkový,

---

[1] Jak zjištěna nezávislost?

[2] Jak **je možné zjistit**, že klient **(ne)bude defaultní, když** mu **neposkytnu půjčku?** U všech „odhadů takové" chyby II. druhu se vypovídá o algoritmu přidělování půjček, nikoliv o charakteru klienta(ů).

průběhem jiné proměnné – autorka připouští závislosti). Pokud by se jednalo o druhou variantu, pak test na přímkovost také nemá smysl (viz např. str. 22, zde uvedené vztahy o „přímkovosti" říkají něco jiného). O co tedy jde?

- Str. 69 a i další. Je uváděn pojem „chyba odhadu koeficientu". Jak je stanoven a za jakých předpokladů je takový odhad konstruován?
- Str. 71. Jaký má význam a co znamená proměnná „délka pracovního poměru" u živnostníka?
- Str. 73. Úspěšnost modelu je zde testována patrně za předpokladu, že chyby I. a II. druhu mají stejné efekty!
- Str. 74 – 79. Poznámky, otázky a připomínky uvedené k modelu „živnostník" lze s malými terminologickými obměnami aplikovat i na model „obchodní společnost".
- Str. 86-87, Závěr. Zde uvedené výroky o „vysoké predikční úspěšnosti" jsou silně nepřesvědčivé. Toto tvrzení je odvozeno z výše uvedených poznámek, otázek a připomínek.

Zásadní náměty do diskuse nad diplomovou prací při její obhajobě

1. Autorka by měla uvést zcela konkrétní případy neparametrických metod, které se „umějí vyrovnat" s multikolinearitou" (viz poznámka pod čarou [18] na str. 18).
2. Co to znamená „perfektně závislé sloupce"? Viz str. 20, předpoklad **D**?
3. Jaké jsou předpoklady pro užití „logitového modelu" (= předpoklady asi jen postačující, pro to, aby pro výraz (3.25) existovalo a bylo jediné globální maximum vůči β). Viz str. 23.
4. Co to je korelace mezi kategoriálními náhodnými proměnnými (str. 49)?
5. Jak to vlastně bylo s testováním algoritmů. Kdy a pro co byl použit soubor pro statistickou inferenci a kdy a pro co byl použit zbývající „testovací, ověřovací" zbytkový soubor.
6. Jak je stanoven odhad „chyby odhadu koeficientu" a za jakých předpokladů je takový odhad konstruován (viz např. str. 69)?

Závěr

Zadání práce, pokud bylo splněno, pak jen v minimálním možném jeho výkladu. Vzhledem k tomu a uvedeným připomínkám **doporučuji** předloženou diplomovou práci k obhajobě před státní zkušební komisí s návrhem klasifikačního stupně zcela určitě ne lepším než **dobře**.

V Plzni dne 21.8.2012

.....Dvořák.....